UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/605,630	10/15/2003	Alain Franciosa	D/A3358	2629

25453 7590 06/04/2007

PATENT DOCUMENTATION CENTER
XEROX CORPORATION
100 CLINTON AVE., SOUTH, XEROX SQUARE, 20TH FLOOR
ROCHESTER, NY 14644

EXAMINER	
SAEED, USMAAN	
ART UNIT	**PAPER NUMBER**
2166	

MAIL DATE	DELIVERY MODE
06/04/2007	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	Application No.	Applicant(s)	
	10/605,630	FRANCIOSA ET AL.	
	Examiner	Art Unit	
	Usmaan Saeed	2166	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE *3* MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *19 March 2007*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1,3,4 and 6-22* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1,3,4 and 6-22* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *15 October 2003* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☐ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

Continued Examination Under 37 CFR 1.114

1. A request for continued examination under 37 CFR 1.114, including the fee set forth in 37 CFR 1.17(e), was filed in this application after final rejection. Since this application is eligible for continued examination under 37 CFR 1.114, and the fee set forth in 37 CFR 1.17(e) has been timely paid, the finality of the previous Office action has been withdrawn pursuant to 37 CFR 1.114. Applicant's submission filed on 03/19/2007 has been entered.

Claim Rejections - 35 USC § 103

2. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

This application currently names joint inventors. In considering patentability of the claims under 35 U.S.C. 103(a), the examiner presumes that the subject matter of the various claims was commonly owned at the time any inventions covered therein were made absent any evidence to the contrary. Applicant is advised of the obligation under 37 CFR 1.56 to point out the inventor and invention dates of each claim that was not

commonly owned at the time a later invention was made in order for the examiner to consider the applicability of 35 U.S.C. 103(c) and potential 35 U.S.C. 102(e), (f) or (g) prior art under 35 U.S.C. 103(a).

Claims 1, 8-13, and 15 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Rie Kubota.** (**Kubota** hereinafter) (U.S. Patent No. 6,041,323) in view of **Gregory Grefenstette** (**Grefenstette** hereinafter) (U.S. Patent No. 6,396,951).

With respect to claim 1, **Kubota** teaches **a method for identifying output documents similar to an input document, comprising:**

"(a) identifying a predefined number of keywords from a first list of rated keywords extracted from the input document to define a list of best keywords; the list of best keywords having a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency" as extracting a partial input character string from the input document, and determining whether the partial input character string is candidate character string (**Kubota** Col 3, Lines 40-42). A unique character string extracted from the input sentence is weighted by the appearance frequency information of the unique character string (**Kubota** Col 3, Lines 16-18). Such a search requires a search key dictionary. In a method performing extraction based on vocabulary information (word dictionary) such as the search key dictionary (**Kubota** Col

1, Lines 51-54). Examiner interprets if the keywords are not present in the dictionary then they don't have a linguistic frequency.

"(b) formulating a query using the list of best keywords and

(c) performing the query to assemble a first set of output documents" as a method for searching for a comparison document, which has character strings similar to a partial input character string existing in an input document. The search is performed on a plurality of documents to be searched (**Kubota** Col 5, Lines 3-7). Then, the documents found by the search are evaluated (**Kubota** Col 11, line 36). Examiner interprets character strings as an input query.

"(d) identifying lists of keywords for each output document in the first set of documents and

(e) computing a measure of similarity between the input document and each output document in the first set of documents" as a method for evaluating similarity between a comparison document and an input document which contains a first unique character string and a second unique character string input in a computer system, said computer being operable to search a comparison document (**Kubota** Col 5, lines 54-58). Calculating the similarity factor of the comparison document from the first appearance frequency value taking the first weight value into account and the second appearance frequency value taking the second weight value into account (**Kubota** Col 6, Lines 7-11).

"(f) defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input

document is greater than a predetermined threshold value; wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency" as rearranging the located document in the order of evaluation (**Kubota** Col 2, Lines 64-65). "Character strings similar to the unique character string" means character strings resembling the unique character string with a predetermined similarity factor or higher, including a character string with a similarity factor of 100%, or complete matching (**Kubota** Col 5, Lines 22-26). Such a search requires a search key dictionary. In a method performing extraction based on vocabulary information (word dictionary) such as the search key dictionary (**Kubota** Col 1, Lines 51-54). The best keywords are less since the dictionary has no errors in its list.

"each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document" as in the case of multiple documents, it may be a set of documents including the input document, or a set of document extracted by search or the like (**Kubota** Col 3, Lines 63-66).

"wherein the query is repeated until a predetermined number of results are obtained or the query is terminated" as inputting a search condition such as AND or OR, or selecting the number of documents set to be extracted as the set of search results, or an allowable similarity factor (**Kubota** Col 12, Lines 62-65). Examiner interprets number of documents set to be extracted as a predetermined number.

"**(g) if the second set of documents includes a matching document but no similar documents repeating (a)-(f) using the matching document to identify similar documents"** as 915 is one for selecting whether the subject of search is an entire document or a set of partial documents such as a set of searched documents. When the search is performed again for the set of searched documents, unique character strings are extracted by comparing the input documents and a set of documents as the result of search limited to a category. Thus, it is possible to extract a character string which is a feature of the input document from a plurality of documents containing similar contents. In addition, the pull-down menu 915 enables selective searching for a limited part of a document such as searching for only titles, instead of the entire document (**Kubota** Col 13, Lines 1-12 and figure 11). In figure 11, reference numeral 947 is performing a similarity search based on the document outputted as a search result.

Kubota teaches the elements of claim 1 as noted above but does not explicitly teaches **"tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages."**

However, **Grefenstette** teaches **"tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages"** as the text code data can be tokenized to obtain token data; the token data can be disambiguated to obtain disambiguated data with parts of speech for words; the disambiguated data can

be lemmatized to obtain lemmatized data indicating, for each of a set of words, either the word or a lemma for the word; and the lemmatized data can be translated. Translation can be done by looking up the words and lemmas in a bilingual translation dictionary (**Grefenstette** Col 2, Lines 19-28).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Grefenstette's** teachings would have allowed **Kubota** to provide automatic translation, by using a bilingual database, parallel corpora, or a manually or automatically constructed bilingual lexicon constructed from parallel corpora to retrieve and display documents in different languages.

With respect to claims 8 and 9 **Kubota** teaches **"the method according to claim 1, further comprising: receiving an input document having textual content and image content; performing OCR on the image content to identify text; analyzing the text and the textual content to identify keywords and recording a digital image representation of the input document; performing OCR on the digital image representation to identify text; analyzing the text to identify keywords."** as in step 404, one document is read from the database 202 to the memory region obtained in step 402. In step 406, the above-mentioned normalization is performed for the document read in step 404. In step 408, fixed length chains, variable length chains, and delimiter patterns are created by scanning the normalized document

(**Kubota** Col 24, Lines 39-44). Contents of individual documents are searchably stored, for example, in a text file form (**Kubota** Col 9, Lines 44-45).

Kubota teaches the elements of claims 8 and 9 but does not explicitly disclose **"performing OCR on the image content to identify text."**

However, **Grefenstette** discloses **"performing OCR on the image content to identify text and recording a digital image representation of the input document"** as automatic recognition can be implemented with optical character recognition (OCR), and automatic language identification can be performed to identify the probable predominant language so that language-specific OCR can be performed. The OCR results can also be presented to the user, who can interactively modify them to obtain the text code data (**Grefenstette** Col 2, Lines 12-18).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Grefenstette's** teachings would have allowed **Kubota** to provide automatic translation, by using a bilingual database, parallel corpora, or a manually or automatically constructed bilingual lexicon constructed from parallel corpora to retrieve and display documents in different languages.

With respect to claim 10, **Kubota** teaches **the method according to claim 1, further comprising:**

"(k) extracting from the input document the first list of keywords" as extracting a partial input character string from the input document, and determining

whether the partial input character string is candidate character string (**Kubota** Col 3, Lines 40-42).

"(l) determining if each keyword in the first list of keywords exists in a domain specific dictionary of words" as a search requires a search key dictionary. In a method performing extraction based on vocabulary information (word dictionary) such as the search key dictionary (**Kubota** Col 1, Lines 51-54).

"(m) for each keyword in the first list of keywords, determining its frequency of occurrence in the input document, also referred to as its term frequency" as a unique character string extracted from the input sentence is weighted by the appearance frequency information of the unique character string (**Kubota** Col 3, Lines 16-18).

"(n) for each keyword identified at (h) that exists in the domain specific dictionary of words, assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies; (o) for each keyword that was not identified in the domain specific dictionary of words at (h), assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; (p) for each keyword in the first list of keywords to which a term frequency and a linguistic frequency are assigned, computing a rating corresponding to its importance in the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence

in the collection of documents" as the following three factors are selectable among the factors to decide the score of document:

a. Frequency of search terms in the document As the search term appears more frequently in the document, the score of the document gets higher.

b. Frequency of search terms in the whole set of documents as the search term appears less frequently in the whole set of documents (all the documents indexed), the search term contributes to the score of the document more.

c. Weight parameter specified explicitly by the user program as the weight of the search term is larger, the search term contributes to the score of the document more (**Kubota** Col 16, Lines 14-28). "Appearance frequency information" means information relating to the number of appearances of a part of the candidate character string in the input document, the comparison document or the like, and may be not only the number of appearances derived by investigating all of a documents, but also information based on the number of appearance in a sample of each document (**Kubota** Col 4, Lines 20-26). The number of appearances may be effected such that 1.5 is added to each appearance of a character string at a position in a document with higher importance such as a heading or title in the input document, while a smaller value of 0.5 is added to the number appearances at a position in a document with less importance such as a footnote or a quotation (**Kubota** Col 15, Lines 53-59). Examiner interprets that if a word does not exist in the dictionary then it does not have a linguistic frequency.

With respect to claim 11, **Kubota** teaches **"the method according to claim 10, for each keyword that was not identified in the domain specific dictionary of words at (l) and that was not assigned at (n) a linguistic frequency from the database of linguistic frequencies, assigning each that matches a regular expression from a set of regular expressions a predefined rating"** as points can be assigned according to Equation (1) in such a manner that (1) a higher point is given to a candidate character string containing an N-character chain with less appearance frequency in the entire set of documents, but higher appearance frequency in the input sentence, and (2) a higher point is given to a candidate character string with a higher appearance frequency in the input sentence (**Kubota** Col 15, Lines 1-9).

With respect to claim 12, **Kubota** teaches **"the method according to claim 11, further comprising, for each keyword in the first list of keywords, modifying the term frequency of keywords determined at (m) to a predefined maximum"** as when the "similarity factor" becomes the maximum value of 1, the character strings completely match. When the character strings completely match, the "similarity factor" always becomes 1 (**Kubota** Col 30, Lines 1-31).

With respect to claim 13, **Kubota** teaches **"the method according to claim 12, wherein keywords include phrases of keywords"** as the search may accommodate new words or phrases, and perform a document search using a request of a user for document search (**Kubota** Abstract).

With respect to claim 15, **Kubota** teaches **"the method according to claim 11, wherein keywords that do not match a regular expression from the set of regular expressions are removed from the first list of keywords"** as If M=2, "communi" is the matched character string. In this case, because of the longest selection, "com" or "commu" is not referred to a matched character string. In addition, "t" is also not a matched character string because it is less than two characters (**Kubota** Col 28, Lines 49-53). Character strings, which divide alphanumeric/katakana are eliminated from the candidate character strings (**Kubota** Col 11, Lines 22-24).

3. Claims 3-4 6-7, 16-21 are rejected under 35 U.S.C. 103(a) as being unpatentable over **Rie Kubota.** (U.S. Patent No. 6,041,323) in view of **Gregory Grefenstette** (U.S. Patent No. 6,396,951) as applied to claims 1, 8-13, and 15 above, further in view of **Gilfillan et al**. (**Gilfillan** hereinafter) U.S. PG Pub No. 2002/0165856.

With respect to claims 3, 4, and 7 **Kubota and Grefenstette** do not explicitly teach **"the method according to claim 2, further comprising (h) if the second set of document contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency, (i): replacing the list of best keywords using keywords having a rating greater than other keywords in the**

first list of rated keywords; and repeating (b)-(f) and the predefined number of keywords identified from the first list of rated keywords is five."

However, **Gilfillan** discloses the systems, which include collaborative research tools to assist with structuring and refining searches over a wide array of disparate data sources. The systems further permit variable access control to research results, for viewing and for editing, throughout iterative stages of research. Research may be conducted with varying degrees of collaboration over varying stages of research refinement, thus providing an end-to-end collaborative research tool that concludes with network publication of organized search results (**Gilfillan** Paragraph 0007).

Further **Gilfillan** teaches if the results are not sufficient, the user may refine the interest as shown in step 510. This may include, for example, removing search terms, adding search terms, replacing search terms, and so forth (**Gilfillan** Paragraph 0060).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Gilfillan's** teachings would have allowed **Kubota and Grefenstette** to provide a platform for sustaining research across available data sources among a number of parties, or over an extended period of time (**Gilfillan** Paragraph 0005) by refining searches and using different search strategies.

Claim 21 is same as claim 4 and is rejected for the same reason as applied hereinabove.

With respect to claim 6, **Kubota** teaches **"the method according to claim 5, performing (i) when textual content in the input document is identified using OCR or a portion of the input document matches the output document"** as in step 404, one document is read from the database 202 to the memory region obtained in step 402. In step 406, the above-mentioned normalization is performed for the document read in step 404. In step 408, fixed length chains, variable length chains, and delimiter patterns are created by scanning the normalized document (**Kubota** Col 24, Lines 39-44). Contents of individual documents are searchably stored, for example, in a text file form (**Kubota** Col 9, Lines 44-45). A method for evaluating similarity between a comparison document and an input document which contains a first unique character string and a second unique character string input in a computer system, said computer being operable to search a comparison document (**Kubota** Col 5, lines 54-58).

With respect to claim 16, **Kubota** teaches **a method for computing ratings of keywords extracted from an input document, comprising:**

"(a) determining if each keyword in the first list of keywords exists in a domain specific dictionary of words" as a search requires a search key dictionary. In a method performing extraction based on vocabulary information (word dictionary) such as the search key dictionary (**Kubota** Col 1, Lines 51-54).

"(b) determining a frequency of occurrence in the input document for each keyword in the list of keywords, also referred to as its term frequency" as a unique

character string extracted from the input sentence is weighted by the appearance frequency information of the unique character string (**Kubota** Col 3, Lines 16-18).

"(c) for each keyword identified at (a) that exists in the domain specific dictionary of words, assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies; (d) for each keyword that was not identified in the domain specific dictionary of words at (a), assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; (e) for each keyword in the first list of keywords to which a term frequency and a linguistic frequency are assigned, computing a rating corresponding to its importance in the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence in the collection of documents" as the following three factors are selectable among the factors to decide the score of document:

a. Frequency of search terms in the document As the search term appears more frequently in the document, the score of the document gets higher.

b. Frequency of search terms in the whole set of documents As the search term appears less frequently in the whole set of documents (all the documents indexed), the search term contributes to the score of the document more.

c. Weight parameter specified explicitly by the user program As the weight of the search term is larger, the search term contributes to the score of the document more (**Kubota**

Col 16, Lines 14-28). "Appearance frequency information" means information relating to the number of appearances of a part of the candidate character string in the input document, the comparison document or the like, and may be not only the number of appearances derived by investigating all of a documents, but also information based on the number of appearance in a sample of each document (**Kubota** Col 4, Lines 20-26).

The number of appearances may be effected such that 1.5 is added to each appearance of a character string at a position in a document with higher importance such as a heading or title in the input document, while a smaller value of 0.5 is added to the number appearances at a position in a document with less importance such as a footnote or a quotation (**Kubota** Col 15, Lines 53-59). Examiner interprets that if a word does not exist in the dictionary then it does not have a linguistic frequency.

"wherein the query is repeated until a predetermined number of results are obtained or the query is terminated" as inputting a search condition such as AND or OR, or selecting the number of documents set to be extracted as the set of search results, or an allowable similarity factor (**Kubota** Col 12, Lines 62-65). Examiner interprets number of documents set to be extracted as a predetermined number.

"**(f) if the second set of documents includes a matching document but no similar documents repeating (a)-(f) using the matching document to identify similar documents**" as 915 is one for selecting whether the subject of search is an entire document or a set of partial documents such as a set of searched documents. When the search is performed again for the set of searched documents, unique character strings are extracted by comparing the input documents and a set of

documents as the result of search limited to a category. Thus, it is possible to extract a character string which is a feature of the input document from a plurality of documents containing similar contents. In addition, the pull-down menu 915 enables selective searching for a limited part of a document such as searching for only titles, instead of the entire document (**Kubota** Col 13, Lines 1-12 and figure 11). In figure 11, reference numeral 947 is performing a similarity search based on the document outputted as a search result.

Kubota teaches the elements of claim 1 as noted above but does not explicitly teaches **"tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages, and wherein a query reduction is performed by removing at least one keyword in the list of best keywords that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency if an insufficient number of results are obtained from the list of keywords."**

However, **Grefenstette** teaches **"tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages"** as the text code data can be tokenized to obtain token data; the token data can be disambiguated to obtain disambiguated data with parts of speech for words; the disambiguated data can be lemmatized to obtain lemmatized data indicating, for each of a set of words, either the word or a lemma for the word; and the lemmatized data can be translated.

Translation can be done by looking up the words and lemmas in a bilingual translation dictionary (**Grefenstette** Col 2, Lines 19-28).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Grefenstette's** teachings would have allowed **Kubota** to provide automatic translation, by using a bilingual database, parallel corpora, or a manually or automatically constructed bilingual lexicon constructed from parallel corpora to retrieve and display documents in different languages.

Kubota and Grefenstette teach the elements of claim 16 as noted above but do not explicitly disclose **"wherein a query reduction is performed by removing at least one keyword in the list of best keywords that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency if an insufficient number of results are obtained from the list of keywords."**

However, **Gilfillan** teaches **"wherein a query reduction is performed by removing at least one keyword in the list of best keywords that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency if an insufficient number of results are obtained from the list of keywords"** as a systems, which include collaborative research tools to assist with structuring and refining searches over a wide array of disparate data sources. The systems further permit variable access control to research results, for viewing and for editing, throughout iterative stages of research. Research may be conducted with varying degrees of collaboration over varying stages of research refinement, thus

providing an end-to-end collaborative research tool that concludes with network publication of organized search results (**Gilfillan** Paragraph 0007).

Further **Gilfillan** teaches if the results are not sufficient, the user may refine the interest as shown in step 510. This may include, for example, removing search terms, adding search terms, replacing search terms, and so forth (**Gilfillan** Paragraph 0060).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Gilfillan's** teachings would have allowed **Kubota and Grefenstette** to provide a platform for sustaining research across available data sources among a number of parties, or over an extended period of time (**Gilfillan** Paragraph 0005) by refining searches and using different search strategies.

With respect to claim 17, **Kubota** teaches **"the method according to claim 16, wherein the keywords in the list of keywords are used to carry out one of language identification, indexing, categorization, clustering, searching, translating, storing, duplicate detection, and filtering"** as if there are multiple documents describing "methods for searching documents for example, there is a high possibility that the keywords being extracted are very similar ones such as "search", "character string", and "high speed" (**Kubota** Col 2, Lines 24-28). Input sentence" described herein means one or more sentences in a language such as Japanese or English (**Kubota** Col 2, Lines 66-67). Unique character strings are extracted by

comparing the input document and a set of documents as the result of search limited to a category (**Kubota** Col 13, Lines 4-7).

Claim 19 is essentially the same as claim 10 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

With respect to claim 20, **Kubota** teaches a**n article of manufacture for identifying output documents similar to an input document, the article of manufacture comprising computer usable media including computer readable instructions embedded therein that causes a computer to perform a method, wherein the method comprises:**

"(a) identifying a predefined number of keywords from a first list of rated keywords extracted from the input document to define a list of best keywords; the list of best keywords having a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency" as extracting a partial input character string from the input document, and determining whether the partial input character string is candidate character string (**Kubota** Col 3, Lines 40-42). A unique character string extracted from the input sentence is weighted by the appearance frequency information of the unique character string (**Kubota** Col 3, Lines 16-18). Such a search requires a search key dictionary. In a method performing extraction based on vocabulary information (word dictionary) such as the search key dictionary (**Kubota** Col

1, Lines 51-54). Examiner interprets if the keywords are not present in the dictionary then they don't have a linguistic frequency.

"(b) formulating a query using the list of best keywords and

(c) performing the query to assemble a first set of output documents" as a method for searching for a comparison document, which has character strings similar to a partial input character string existing in an input document. The search is performed on a plurality of documents to be searched (**Kubota** Col 5, Lines 3-7). Then, the documents found by the search are evaluated (**Kubota** Col 11, line 36). Examiner interprets character strings as an input query.

"(d) identifying lists of keywords for each output document in the first set of documents and

(e) computing a measure of similarity between the input document and each output document in the first set of documents" as a method for evaluating similarity between a comparison document and an input document which contains a first unique character string and a second unique character string input in a computer system, said computer being operable to search a comparison document (**Kubota** Col 5, lines 54-58). Calculating the similarity factor of the comparison document from the first appearance frequency value taking the first weight value into account and the second appearance frequency value taking the second weight value into account (**Kubota** Col 6, Lines 7-11).

"(f) defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input

document is greater than a predetermined threshold value; wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency" as rearranging the located document in the order of evaluation (**Kubota** Col 2, Lines 64-65). "Character strings similar to the unique character string" means character strings resembling the unique character string with a predetermined similarity factor or higher, including a character string with a similarity factor of 100%, or complete matching (**Kubota** Col 5, Lines 22-26). Such a search requires a search key dictionary. In a method performing extraction based on vocabulary information (word dictionary) such as the search key dictionary (**Kubota** Col 1, Lines 51-54). The best keywords are less since the dictionary has no errors in its list.

"each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document" as in the case of multiple documents, it may be a set of documents including the input document, or a set of document extracted by search or the like (**Kubota** Col 3, Lines 63-66).

"wherein the query is repeated until a predetermined number of results are obtained or the query is terminated" as inputting a search condition such as AND or OR, or selecting the number of documents set to be extracted as the set of search results, or an allowable similarity factor (**Kubota** Col 12, Lines 62-65). Examiner interprets number of documents set to be extracted as a predetermined number.

"**(f) if the second set of documents includes a matching document but no similar documents repeating (a)-(f) using the matching document to identify similar documents"** as 915 is one for selecting whether the subject of search is an entire document or a set of partial documents such as a set of searched documents. When the search is performed again for the set of searched documents, unique character strings are extracted by comparing the input documents and a set of documents as the result of search limited to a category. Thus, it is possible to extract a character string which is a feature of the input document from a plurality of documents containing similar contents. In addition, the pull-down menu 915 enables selective searching for a limited part of a document such as searching for only titles, instead of the entire document (**Kubota** Col 13, Lines 1-12 and figure 11). In figure 11, reference numeral 947 is performing a similarity search based on the document outputted as a search result.

Kubota teaches the elements of claim 20 as noted above but does not explicitly teaches **"tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages, and (g) if the second set of documents contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency."**

However, **Grefenstette** teaches **"tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages"** as the text code data can be tokenized to obtain token data; the token data can be disambiguated to obtain disambiguated data with parts of speech for words; the disambiguated data can be lemmatized to obtain lemmatized data indicating, for each of a set of words, either the word or a lemma for the word; and the lemmatized data can be translated. Translation can be done by looking up the words and lemmas in a bilingual translation dictionary (**Grefenstette** Col 2, Lines 19-28).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Grefenstette's** teachings would have allowed **Kubota** to provide automatic translation, by using a bilingual database, parallel corpora, or a manually or automatically constructed bilingual lexicon constructed from parallel corpora to retrieve and display documents in different languages.

Kubota and Grefenstette teach the elements of claim 20 as noted above but do not explicitly discloses **"(g) if the second set of documents contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency."**

However, **Gilfillan** teaches **"(g) if the second set of documents contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency"** as a systems, which include collaborative research tools to assist with structuring and refining searches over a wide array of disparate data sources. The systems further permit variable access control to research results, for viewing and for editing, throughout iterative stages of research. Research may be conducted with varying degrees of collaboration over varying stages of research refinement, thus providing an end-to-end collaborative research tool that concludes with network publication of organized search results (**Gilfillan** Paragraph 0007).

Further **Gilfillan** teaches if the results are not sufficient, the user may refine the interest as shown in step 510. This may include, for example, removing search terms, adding search terms, replacing search terms, and so forth (**Gilfillan** Paragraph 0060).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Gilfillan's** teachings would have allowed **Kubota and Grefenstette** to provide a platform for sustaining research across available data sources among a number of parties, or over an extended period of time (**Gilfillan** Paragraph 0005) by refining searches and using different search strategies.

Claim 18 is essentially the same as claim 20 except it sets forth the claimed invention as a system and is rejected for the same reasons as applied hereinabove.

4. Claims 14 and 22 is rejected under 35 U.S.C. 103(a) as being unpatentable over **Rie Kubota.** (U.S. Patent No. 6,041,323) in view of **Gregory Grefenstette** (U.S. Patent No. 6,396,951) as applied to claims 1, 8-13, and 15 above, in view of **Cofino et al**. (**Cofino** hereinafter) (U.S. PG Pub No. 2005/0187931).

With respect to claims 14 and 22, **Kubota and Grefenstette** do not explicitly teach **"the method according to claim 11, wherein the rating is a weight computed using the following equation: W.sub.t,dF.sub.t,d*log(N/F.sub.t), where: W.sub.t,d: the weight of term tin document d; F.sub.t,d: the frequency occurrence of term tin document d; N: the number of documents in the collection of documents; F.sub.t: the document linguistic frequency of term t in the collection of documents."**

However, **Cofino** discloses **"the method according to claim 11, wherein the rating is a weight computed using the following equation: W.sub.t,dF.sub.t,d*log(N/F.sub.t), where: W.sub.t,d: the weight of term tin document d; F.sub.t,d: the frequency occurrence of term tin document d; N: the number of documents in the collection of documents; F.sub.t: the document linguistic frequency of term t in the collection of documents"** as the most traditional tf.times.idf term weighting is f*log (N/n), where f is the frequency of the word

in the current document, N is the total documents in the local corpus, and n is the number of documents in the local corpus containing the word (**Cofino** Paragraph 0009).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Cofino's** teachings would have allowed **Kubota and Grefenstette** to evaluate the importance of terms and phrases in a document in a personal corpus relative to usage in one or more larger reference corpuses (**Cofino** Paragraph 0013).

Response to Arguments

5. Applicant's arguments have been considered but are moot in view of the new ground(s) of rejection.

In these arguments applicant relies on the amended claims and not the original ones.

Applicant argues that **Kubota** does not teach **"wherein the query is repeated until a predetermined number of results are obtained or the query is terminated," "if the second set of documents includes a matching document but no similar documents repeating (a)-(f) using the matching document to identify similar documents," and "tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages."**

In response to the preceding arguments applicant respectfully submits that **Kubota** teaches **"wherein the query is repeated until a predetermined number of results are obtained or the query is terminated"** as inputting a search condition such as AND or OR, or selecting the number of documents set to be extracted as the set of search results, or an allowable similarity factor (**Kubota** Col 12, Lines 62-65).

Examiner interprets number of documents set to be extracted as a predetermined number.

"if the second set of documents includes a matching document but no similar documents repeating (a)-(f) using the matching document to identify similar documents" as 915 is one for selecting whether the subject of search is an entire document or a set of partial documents such as a set of searched documents. When the search is performed again for the set of searched documents, unique character strings are extracted by comparing the input documents and a set of documents as the result of search limited to a category. Thus, it is possible to extract a character string which is a feature of the input document from a plurality of documents containing similar contents. In addition, the pull-down menu 915 enables selective searching for a limited part of a document such as searching for only titles, instead of the entire document (**Kubota** Col 13, Lines 1-12 and figure 11). In figure 11, reference numeral 947 is performing a similarity search based on the document outputted as a search result to further identify similar documents.

Kubota teaches the elements as noted above but does not explicitly teaches **"tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages."**

However, **Grefenstette** teaches **"tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages"** as the text code data can be tokenized to obtain token data; the token data can be disambiguated to obtain disambiguated data with parts of speech for words; the disambiguated data can be lemmatized to obtain lemmatized data indicating, for each of a set of words, either the word or a lemma for the word; and the lemmatized data can be translated. Translation can be done by looking up the words and lemmas in a bilingual translation dictionary (**Grefenstette** Col 2, Lines 19-28).

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teaching of the cited references because **Grefenstette's** teachings would have allowed **Kubota** to provide automatic translation, by using a bilingual database, parallel corpora, or a manually or automatically constructed bilingual lexicon constructed from parallel corpora to retrieve and display documents in different languages.

Conclusion

6. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure is listed on 892 form.

Examiner's Note: Examiner has cited particular figures, columns and line numbers in the references as applied to the claims above for the convenience of the applicant. Although the specified citations are representative of the teachings in the art and are applied to the specific limitations within the individual claim, other passages and figures may apply as well. It is respectfully requested from the applicant, in preparing the responses, to fully consider the references in entirety as potentially teaching all or part of the claimed invention, as well as the context of the passage as taught by the prior art or disclosed by the examiner.

Contact Information

7. Any inquiry concerning this communication or earlier communications from the examiner should be directed to Usmaan Saeed whose telephone number is (571)272-4046. The examiner can normally be reached on M-F 8-5.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Hosain Alam can be reached on (571)272-3978. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

Usmaan Saeed
Patent Examiner
Art Unit: 2166

Leslie Wong LW
Primary Examiner

US
May 16, 2007

HOSAIN ALAM
SUPERVISORY PATENT EXAMINER